# علم الأصوات والتهجئة العربية
# Arabic Phonology and Orthography

في هذا الفصل نعرض وصفاً مختصراً لعلم أصوات اللغة العربية المعاصرة (فنولوجي Phonology) والمفاهيم المرتبطة بها. متبوعة بوصف لكيفية التهجئة العربية (أي قواعدها الإملائية) في عملية الربط التبادلي بين الصوت والخط العربي. بعدها، نستعرض أربع عمليات حاسوبية ذات علاقة بهذا المجال وهي: النقل الكتابي لأسماء الأعلام (proper name transliteration)، والتصحيح الإملائي (spelling correction)، والتعرف على الكلام (speech recognition) وتوليد الكلام آلياً (speech synthesis) [1].

## ٣,١ علم الأصوات العربية

يقدم هذا القسم عرضاً موجزاً لعلم الأصوات العربية (Arabic phonology) أو ما يطلق عليها الفنولوجيا. وسنقدم تعريفاً للمصطلحات الفنولوجية إذا دعت الحاجة لذلك. ولمزيد من النقاش حول الفنولوجيا في الوضع الحوسبي، يمكن الاطلاع على [36]. وعلى الرغم من أن التركيز في هذا الكتاب سيكون على اللغة العربية المعاصرة، فإن هذا الفصل سيستعرض بعضاً من مسائل اللهجات، وذلك لكون اللهجات في الأساس يتحدث بها، وأحياناً تؤثر في طريقة نطق اللغة العربية المعاصرة. أما العربية المستخدمة في القرآن الكريم، فلن نتطرق لها لكونها تملك قواعدها الخاصة في النطق والإملاء وهي قواعد تختلف من نواحٍ عدة عن اللغة العربية المعاصرة وعن اللهجات العربية.

---

(١) أيضاً يطلق عليه توليف الكلام أو تركيبه (توضيح من المترجمة).

### ٣,١,١ مفاهيم أساسية

علم الأصوات (الفنولوجي) هو دراسة لكيفية تنظيم الأصوات في اللغات الطبيعية [37] . والمفهوم الرئيسي في علم الأصوات هو الفونيم (phoneme)[١] ، الذي يمثل أصغر وحدة مقابلة في النظام الصوتي للغة. ويقصد بالمقابلة هنا هو أن اللغة المعنية لديها حد أدنى من الأزواج التي تتضمن الفونيم، بمعنى آخر: توجد كلمتان بمعان مختلفة ويحدث أن يكون الاختلاف فنولوجياً في ذلك الفونيم فقط. على سبيل المثال، كلمتي قلب /qalb/ وكلب /kalb/ في اللغة العربية المعاصرة، يمثلان الحد الأدنى من الأزواج للفونيمان: ق /q/ و ك /k/.

يمكن للفونيم أن يتوافق مع عدة أصوات أو ألوفونات (phones)، أو أصوات أساسية، توزع وفقاً لقواعد يمكن التنبؤ بها تسمى التكتيك الصوتي (phonotactics)[٢]. وتسمى الأصوات التي يمكن التنبؤ بها والمرتبطة مع الفونيم بالألوفونات (allophones)[٣]، بمعنى آخر هي مجموعة الصور المختلفة للفونيم التي تظهر في جميع السياقات الصوتية المحتملة. على سبيل المثال، في حين أن اللغة العربية لا تملك فونيم /p/ مما تسبب في كثير من الأحيان في اللبس الذي يتميز به الناطقون باللهجات العربية في نطق الحرفان p-b ، ويظهر الصوت [p] كالألفون (الصوت) من الفونيم /b/ في سياقات محدودة، كأن يسبقها صوت مهموس، مثال ذلك: كلمة دبس (dibs) تمثل

(١) ويطلق عليها (الوحدة الصوتية): وهي أصغر وحدة صوتية مجردة في اللغة (توضيح من المترجمة).

(٢) علم يختص بتراكب الأصوات المحددة في كل لغة، بمعنى أن كل لغة لها تتابع أصوات لا يتغير (توضيح من المترجمة).

(٣) اللوين الصوتي أو المتغير الصوتي (allophone) هو "صوت كلامي حقيقي يتوزع بطريقة تكاملية أو يتغير بشكل حر" (توضيح من المترجمة).

صوتميا (phonemically) بـ /dibs/ أما فنولوجيا (phonetically) بشكل [dips][1] .

**تنبيه مصـــطلحي**: استخدام مصطلحات مثل فونيم وتكتيك صوتي في مجال المعالجة الطبيعية للغة (NLP) مثل التعرف الآلي على الكلام قد لا يكون متوافقاً تماماً مع كيفية استخدامها عند اللغويين. على سبيل المثال، قد يشير التكتيك الصوتي في مجال المعالجة الآلية للغة إلى سلسة إن قرام (n-gram) للأصوات/الفونيميات، بدلاً من الإشارة إلى قواعد لغوية واضحة.

## ٣,١,٢ مخطط لعلم الأصوات العربي

**الشكل رقم (٣,١). جدول الفونيمات للصوامت العربية. تمثل الصفوف طرق النطق المختلفة، بينما تمثل الأعمدة مختلف مواضع النطق. تم بيان أزواج الفونيمات كمتغيرات واضحة ومفخمة. بينما تعني الفونيمات باللون الرمادي أصوات اللهجات.**

| | | شفوي Labial | شفوي أسناني Labio-dental | بين أسناني Interdental | أسناني Dental | لثوي Alveolar | غاري Palatal | طبقي Velar | لهوي Uvular | حلقي Pharyngeal | حنجري Glottal |
|---|---|---|---|---|---|---|---|---|---|---|---|
| وقف Stop | مهموس voiceless | | | | t ت T ط | | | k ك | q ق | | |
| | مجهور Voiced | b ب | | | d د D ض | | | g | Q | H ح | h ه |
| رخو Fricative | مهموس voiceless | | f ف | θ ث | s س S ص | š ش | | x خ | | ʕ ع | |
| | مجهور Voiced | | | ð ذ Ď ظ | z ز Z | ž | | ɣ | | | |

(١) في هذا الكتاب سنستخدم الحواصر المائلة /..../ للدلالة على التسلسل الصوتمي (phonemic) والأقواس المربعة [...] للدلالة على التسلسل الصوتي (phonetic). كما سنستخدم النقل الكتابي (لحبش وسودي وبكوالتر) [٤] مع بعض الإضافات عوضاً عن رموز المنظمة العالمية للصوتيات ( International Phonetic Alphabet) للتقليل من عدد التمثيلات المستخدمة في هذا الكتاب.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | č | | | | | مهموس voiceless | مزجي Affricate |
| | | | | | ج j | | | | | مجهور Voiced | |
| | | | | ي y | | | | | و w | | انزلاقي Glide |
| | | | | | ن n | | | | م m | | غني Nasal |
| | | | | | ر r | | | | | | سائل Liquid |

**الشكل رقم (٣,٢). جدول الفونيمات للصوائت العربية. تمثل الصوائت بارتفاع وارتداد موضع اللسان. أما الفونيمات باللون الرمادي فتعني أصوات اللهجات.**

| مؤخرة اللسان Back | وسط اللسان Central | مقدمة اللسان Front | |
|---|---|---|---|
| u ū | | i ī | مرتفع High |
| o ō | | e ē | متوسط Mid |
| | a ā | | منخفض Low |

تتضمن القائمة الفنولوجية الأساسية في اللغة العربية المعاصرة ٢٨ صامتاً (consonants) وثلاثة صوائت قصيرة (short vowels) وثلاثة صوائت طويلة ( long vowels)، بالإضافة إلى صائتين مركبين (إدغامين (diphthongs)) هما /ay/ و/aw/ . يظهر الشكلان رقما ٣,١ و ٣,٢ مختلف الصوامت والصوائت الفونيمية في اللغة العربية المعاصرة بناء على خصائص النطق (وذلك بعد الرجوع إلى مصدر [39, 38]).

في الشكل رقم (٣,١)، يدل وجود زوج من الفونيمات في خانة واحدة، كما في ' t T' على أنهما واضحان ومؤكدان عند النطق. يعرف التفخيم (Emphasis) على أنه تأثير باس (bass effect)[1] الذي يعطي انطباعاً صوتياً بصدى أجوف للصوت

(١) يرمز مصطلح باس (bass) للتأثير الصوتي العميق (توضيح من المترجمة).

الأساسي [38] . وهذا النوع من التفخيم الحلقي، أي التفخيم مع وجود ثمانية من الصوامت في منطقة الحلق وآخره، هو ما يميز النطق باللغة العربية [38] . أزواج الفونيمات الصوائتية (Vowel phoneme pairs) في الشكل رقم ٣,٢ تبين الاختلاف في الطول (بين الصوائت الطويلة والقصيرة). ولا تعتبر الفونيمات باللون الرمادي في الجدول رقم (٣,١ و ٣,٢) من اللغة العربية المعاصرة، بل من اللهجات. وللمزيد من المعلومات حول هذا الموضوع انظر في قسم ٣,١,٣.

جميع الصوامت في العربية لديها مقابل في الإنجليزية مع وجود الاستثناءات التالية [١]:

- /H/ صوته كأنه h مع وجود هسهسة يمكن تقريب تشبيهها بالصوت الصادر عند التنفس على النظارات قبل مسحها لتنظيفها.
- /ç/ هو البديل المنطوق من /H/ الذي يبدو كصوت /a/ حاد.
- /x/ تماثل loch في الأسكتلندية أو chutzpa في الإنجليزية - اليهودية (Yiddish-English).
- /r/ و /γ/ تعادل r في الإسبانية و r الفارسية- الفرنسية.
- الوقف اللهوي /q/ صوته كأنه باس عميق /k/.
- الهمزة /ʾ/ تبدو كأنها الصوت الوسطي في الكلمة الإنجليزية uh-oh.
- الأصوات المفخمة ( /D/و /T/ و /S/ و /Ď/ ) أعطيت جودة باس (bass quality) لمقابلاتها العادية ((/d/ /t/ /s/ /ð/ )).

(١) لم نُضمن فونيمات مفخمة إضافية تظهر في عدد محدود من الأزواج القليلة مثل الحرف المفخم /l/ و /b/ أو الفونيمات المستعارة من كلمات للغات أجنبية مثل /v/ و /p/.

لاحظ أن الفونيمات /š/ و /θ/ تعادل الفونيمات الإنجليزية sh و th التي تكتب كحرفين مركبين.

الفونيمات الصوائتية في اللغة العربية المعاصرة محدودة العدد مقارنة بالإنجليزية والفرنسية، إلا أن هناك الكثير من الألفونات لكل واحد منها، وهذه الألفونات تعتمد على سياق الصوامت [38] . على سبيل المثال، قارن نطق الصائت / ā / في كلمة باس / bās / مع كلمة باص /bāS/ التي يمكن مقاربتها مع الكلمة الإنجليزية (bass)- وهو نوع من أنواع الأسماك وكلمة (boss)- التي تعني الرئيس- . وتسمى هذه الظاهرة انتشار التفخيم. وهو تكتيك صوتي شائع بحيث تصبح الصوائت والصوامت القريبة من الصامت المفخم ممثلة بألوفوناتها المفخمة. كما أن هناك ظاهرة أخرى مثيرة للاهتمام في نطق صوائت اللغة العربية المعاصرة وهي اختيارية حذف الصائت الأخير الذي يبين حالتها النحوية في نهاية لفظ الكلمة (كما هو الحال في نهاية الجملة أو نهاية الاقتباس). وهذا ما يسمى بالوقف.

هناك العديد من التنوعات الفنولوجية الإضافية المقتصرة على سياق صرفي محدد، بمعنى أنها مقيدة صرف- صوتياً مقارنة بفنولوجيا. وتمثل بعضاً من هذه الظواهر هجائياً بشكل صريح وبعضها لا يمثل. نطلق على الحالات التي تمثل هجائيا بالتعديلات الصرفية (morphological adjustments) وسنناقشها في الفصل القادم. على سبيل المثال، الفونيم /t/ في النمط الفعلي افتعل (VIII) تصبح مجهورة ويمكن تهجئتها (ليس فقط نطقها) كـ "د" "d"عندما تكون ملاصقة لصوامت جذر محدد. في المقابل، نسمى الحالات التي لا يمكن تمثيلها هجائياً، مثل الإدغام الفونولوجي ( phonological assimilation) لـ ال التعريف ( ال+ + Al) لبعض الفونيمات التي تتبعها، بالإملاء الصرف- فونيمي (morpho-phonemic spelling). سنناقش هذه الحالات في قسم

٣,٢,٣. ولنقاش موسع حول فونولوجيا اللغة العربية المعاصرة والنبر (stress) وبنيتها المقطعية (syllabic structure) انظر [39.38] .

### ٣,١,٣ الاختلافات الفونولوجية بين اللهجات العربية واللغة العربية المعاصرة

تختلف اللهجات العربية فونولوجياً عن اللغة العربية المعاصرة وأيضاً يختلف بعضها عن بعض. وبعض من الاختلافات الشائعة يتضمن ما يلي [38 . 40 . 41] :

- في اللغة العربية المعاصرة الأحرف اللثوية الانفجارية أو ما تسمى بالفيولر أفريكت (alveolar affricate) مثل ج /j/ يتم نطقها /g/ بالمصرية و /ž/ بالشامية و /y/ بالخليجية. على سبيل المثال، جميل تنطق /jamīl/ بالعربية المعاصرة والعراقية، و/gamīl/ بالمصرية، و/žamīl/ بالشامية، و/yamīl/ بالخليجية. ويعتبر النطق الشامي والمصري هو المعيار للغة العربية المعاصرة في هذه المناطق[١].
- في اللغة العربية المعاصرة الصامت ق /q/ يتم تمييزها كهمزة /'/ في مصر والشام وك /g/ في الخليج والعراق. على سبيل المثال كلمة طريق تظهر /Tarīq/ في العربية المعاصرة و/Tarī'/ في المصرية والشامية و/Tarīg/ في العراقية والخليجية. كما توجد اختلافات أخرى موجودة في بعض اللهجات الفرعية مثل نطق /k/ في الريف الفلسطيني (لهجة شامية)، و /j/ في الإماراتية (لهجة خليجية) و /Q/ (/q/ المجهور) في السودانية (لهجة مصرية). هذه التغييرات لا تنطبق على الاستعارات الحديثة والدينية من اللغة العربية المعاصرة. على سبيل المثال، قرآن 'Quran' لا ينطق إلا /qur'ān/.

---

١ نطق حرف الجيم يختلف باختلاف البلدان العربية، فللجيم ثلاثة أنواع من النطق لكل نوع مرجعه التاريخي، انظر ورقة الدكتور خليل محمود عساكر. (١٩٥٥م) طريقة لكتابة نصوص اللهجات العربية الحديثة بحروف عربية. مجلة المجمع اللغوي بالقاهرة. م٨، ص ص ١٨١- ١٩٢ (تعليق من المترجمة).

- في اللغة العربية المعاصرة، يميز الصامت (ك /k/) على أنه /k/ في اللهجات العربية باستثناء الخليج، أما العراقية واللهجة الفرعية للريف الفلسطيني للشامية، فتسمح بنطقها /č/ في بعض السياقات. على سبيل المثال، كلمة سمك تنطق /samak/ في العربية المعاصرة والمصرية ومعظم اللهجة الشامية لكن في العراقية والخليجية تنطق /simač/.
- في اللغة العربية المعاصرة الصامت ث /θ/ ينطق على أنه /t/ في الشامية والمصرية (أو /s/ في استعارات حديثة من العربية المعاصرة)، مثال ذلك، كلمة ثلاثة تنطق /θalāθa/ في العربية المعاصرة مقارنة بـ/talāta/ في المصرية.
- في اللغة العربية المعاصرة الصامت ذ /ð/ ينطق على أنه /d/ في الشامية والمصرية (أو /z/ في استعارات حديثة من العربية المعاصرة)، مثال ذلك، كلمة هذا التي تنطق /hāða/ في العربية المعاصرة مقارنة بـ /hāda/ في الشامية و/da/ في المصرية.
- في اللغة العربية المعاصرة الصامت ض /D/ (d مفخمة) و ظ /Ď/ (/ð/ مفخمة) يتم تسويتهما إلى /D/ في المصرية والشامية و لـ /Ď/ في الخليجية والعراقية.على سبيل المثال، جملة "ظل يضرب" تنطق /Ďalla yaDrubu/ في العربية المعاصرة مقارنة بـ /Dall yuDrub/ في الشامية و/Ďall yuĎrub/ في الخليجية.

  في الاستعارات الحديثة من العربية المعاصرة، تنطق /Ď/ على أنها /Z/ (z مفخمة) في المصرية والشامية. على سبيل المثال، كلمة "ظابط" تنطق /ĎābiT/ في العربية المعاصرة وتنطق /ZābiT/ في المصرية والشامية.
- التغيرات في حروف العلة تتضمن: (أ) تغييراً أو حذفاً بالكامل للصوائت القصيرة: مثلا كلمة يكتب تنطق /yaktubu/ في العربية المعصرة مقارنة بـ/yiktib/ في العراقية والمصرية أو /yoktob/ في الشامية، (ب) عدم إطالة المد في الصوائت

الطويلة النهائية غير المشددة في بعض اللهجات: مثال كلمة مطارات تنطق /maTārāt/ في العربية المعاصرة مقارنة بـ/maTarāt/ في الشامية والمصرية، (ج) الإدغامات /aw/ و/ay/ في العربية المعاصرة أصبحت في معظمها /ō/ و /ē/ في بعض اللهجات، مثال ذلك: كلمة بيت تنطق /bayt/ في العربية المعاصرة لكن تنطق /bēt/ في المصرية والشامية.

- في بعض اللهجات، يظهر فقدان التأكيد في بعض صوامت اللغة العربية المعاصرة، مثال ذلك، كلمة لطيف تنطق /laṬīf/ في العربية المعاصرة في مقابل /latīf/ في لهجة فرعية للشامية لأحد مدن لبنان.

### ٣,٢ التهجئة العربية

علم التهجئة (orthography) هو تحديد كيفية تطابق أصوات اللغة من وإلى خط معين. في هذا القسم، سنقدم سرداً للتهجئة في اللغة المعيارية للعربية المعاصرة باستخدام الخط العربي. ينصب التقابل بين الكتابة والنطق في العربية المعاصرة بين طرق بعض اللغات في هذا المجال كاللغة الإسبانية والفنلندية التي تطابق كل حرف مع صوته، بينما في لغات مثل الإنجليزية والفرنسية، يكون تطابق الحرف مع صوته أكثر تعقيداً [42].

في اللغة العربية المعاصرة، يوجد هناك ٣٤ فونيماً (٢٨ منها للصوامت و ٣ للصوائت الطويلة و٣ للصوائت القصيرة). ولدى الخط العربي ٣٦ حرفاً و ٩ حركات (بما فيها الألف الخنجرية). ومعظم الحروف العربية لديها تطابق (واحد لواحد) مع فونيمات العربية المعاصرة (انظر الشكل رقم ٣,١). إلا أن هناك استثناءات مهمة [42, 4] سنخلصها في التالي.


**الشكل رقم (٣,٣). مقابلة كل حرف عربي مع صوته.**

### ٣,٢,١ التشكيل الاختياري

التشكيل في الخط العربي يتقابل مع الأصوات التالية:

- الصوائت الثلاثة القصيرة، ـَ *a*، ـُ *u*، ـِ *i*، تمثل الصوائت /a/ , /u/ ,/i/ على التوالي. تستخدم الصوائت القصيرة ـُ u و ـِ i مع الصوامت و w وَ ي y للدلالة على الصوائت الطويلة /ū/ (كما في *uw*) وَ /ī/ (كما في *iy*). يكتب الصائت الطويل /ā/ في الغالب كمجموعة من حركة الصائت القصيرَ a وحرف ا A[1]. وهذا ما يجعل هذه الحروف الثلاثة غامضة[2].
- تمثل حركات التنوين الثلاث ـً ã و ـٌ ũ و ـٍ Ĩ مزيجاً من الصوائت القصيرة وعلامة اسم النكرة /n/ في العربية المعاصرة: /an/ و /un/ و /in/ على التوالي.
- الصامت المستخدم للتطويل علامة الشدة ـّ ~ يكرر أو يطيل الصامت السابق له، مثال ذلك: كتّب *kat~ab* تنطق /kattab/.

---

١ بعض مصادر المعالجة الطبيعية للغة العربية وأبرزها محلل بكوالتر الصرفي [٢٣]، يقوم بإسقاط الفتحة a قبل الحرف A.

٢ انظر الشكل رقم ٣,٣ فلكل حرف صامت يقابله رسم واحد للحرف إلا الحرفين (و) وَ (ي) حيث يمكن أن تقابل صامتاً أو صائتاً. أيضاً انظر كتاب (A Reference Grammar of Modern Standard Arabic) ص ٣٠٢ للمؤلف Karin C. Ryding لمزيد من التوضيح (تعليق من المترجمة).

- علامة السكون ـْ. تبين عدم وجود صائت.
- الصائت الطويل، الألف الخنجرية ـ ا á تمثل الصائت الطويل /ā/ في عدد قليل من الكلمات.

تظهر علامات التشكيل العربية بعد الحرف فقط. وعلى هذا النحو، فإن الكلمات التي تبدأ بصائت قصير تمثل بألف ساكنة إضافية، وتسمى هذه ألف الوصل أو همزة الوصل آ Ä (في الغالب تكتب ا A). همزة الوصل التي تأتي في بداية الكلام أو الجملة تنطق كهمزة قطع مسبوقة بصائت قصير، إلا أن همزة الوصل في وسط الجملة تكون صامتة. على سبيل المثال، جملة (انكتبَ كتابٌ) *Ainkataba kitAbũ* تنطق /inkataba kitābun/' لكن جملة (كتابٌ انكتبَ) *kitAbũ Ainkataba* تنطق /kitābun inkataba/. الهمزة الحقيقية دائماً تنطق همزة قطع. تظهر همزة الوصل في الغالب مع الألف في ال التعريف *Al*. كما يظهر أيضاً في كلمات محددة وتصنيفات مثل الضمائر المتصلة، مثال ذلك "الذي" *Alðy* وفي الأفعال في وزن افتعل أي شكل (VII) (انظر الفصل الرابع).

من أكبر المشاكل في التشكيل هو اختيارية استخدامها. قد لا تكون هناك مشكلة كبيرة عند المطابقة من الصوت إلى الخط ولكن المشكلة تكمن في عكس ذلك. ينحصر استخدام التشكيل في النصوص الدينية والكتب العربية بالمدارس. وفي نصوص أخرى، نجد أن حوالي ١,٥٪ من الكلمات تحتوي على تشكيل. بعض التشكيلات معجمية (lexical) (عندما يختلف معنى الكلمة) وبعضها الآخر تصريفية (inflectional) (عندما تختلف حالة الاسم أو صيغة الفعل). عادة ما يكون التشكيل التصريفي في نهاية الكلمة. وبما أن الحالة الاسمية وصيغة الفعل والتنوين قد اختفت جميعها في اللهجات العربية المنطوقة، فإن الناطقين بالعربية لا ينطقون دوماً هذه التصريفات بشكل صحيح

أو حتى لا ينطقونها إطلاقاً. إلا أن الاستثناءات الملحوظة تظهر في التعابير الرسمية المتكررة مثل جملة السلام عليكم *AlslAm ςlykm* التي تنطق /'assalāmu ςalaykum/.

### ٣,٢,٢ إملاء الهمزة

كما ناقشنا في آخر الفصل السابق، للهمزة /'/ عدة أشكال في الخط العربي هي: ء ' ، آ Ā ، أ Â ، ؤ ŵ ، إ Ă ، ئ ŷ . وهناك قواعد معقدة لإملاء الهمزة التي تعتمد أساساً على سياق حروف العلة والسياق الصرفي [٨]. على سبيل المثال، خذ في عين الاعتبار الأشكال المختلفة للهمزة في معنى الكلمة التالية عندما تتغير حالة علامته كالتالي: بهاء *bahA'ahu* تنطق /bahā'ahu/ (حالة النصب)، بهاؤه *bahAŵuhu* تنطق /bahā'uhu/ (حالة الرفع) وبهائه *bahAŷihi* تنطق /bahā'ihi/ (حالة الجر).

كما أن إملاء الهمزة يزداد تعقيداً عندما يقوم الكتّاب باللغة العربية، وفي كثير من الأحيان، باستبدال الحروف المهموزة بحروف غير مهموزة، مثال ذلك أ Â ⇔ ا *A*، أو من خلال إملاء حرفين، مثال ذلك، ئ ŷ ⇔ ىء *'ý*. في الغالب، هذه الاختلافات الشائعة لا تضفي غموضاً، خاصة عندما تكون في أول الجذع (stem-initial)، مثال ذلك اول/أول *Âwl/Awl*. إلا أنها تتسم بالغموض عندما تكون في وسط الجذع (stem-medial) أو نهاية الجذع (stem-final)، حيث تتجاهل في الغالب، مثال ذلك بدا *bdA* وبدأ *bdÂ*. وقد لوحظ أن الهمزة في بداية الجذع للألف المهموزة عادة ما ينظر إليها من قبل الكتّاب العرب على أنها علامة تشكيل واختيارية مقارنة بحالتي وسط الجذع ونهاية الجذع، التي تعتبر أكثر وجوباً [٥].

### ٣,٢,٣ الإملاء الصرف–صوتي

يحتوي الخط العربي على عدد قليل من الإملاء المعجمي/المورفيمي (morphemic/lexical)، بعض منها شائع جداً مثل:

• **التاء المربوطة**: من المعروف أن التاء المربوطة (ة ħ) عادة ما تكون في نهاية المؤنث. وتظهر فقط في نهاية الكلمة ويتبعها التشكيل فقط. ففي اللغة العربية المعاصرة، تنطق /t/ إلا إذا لم يتبعها صائت (كما في الوقف)، في هذه الحالة تكون صامتة. على سبيل المثال، كلمة المكتبة *Almaktabaħ* تنطق /’almaktabatu/ (في الحالة العادية) أو /’almaktaba/ (في حالة الوقف).

• **الألف المقصورة**: الألف المقصورة (ى ý) عبارة عن علامة صامتة مشتقة، تتبع دائماً الصائت القصير /a/ في نهاية الكلمة. على سبيل المثال، كلتا الكلمتين عصى *ʿaSay* وعصا *ʿaSaA* تنطق /ʿaSa/[١].

• **أل التعريف**: تعتبر أل التعريف زائدة في بداية الكلمة وتدغم مع أول صامت في الاسم أو الصفة التي يغيرها إذا كان الصامت لثوياً (Alveolar) أو أسنانياً (dental) أو بين أسناني (inter-dental) (ماعدا /j/)[٢]. وتسمى مجموعة الصوامت الأربعة عشر بالحروف الشمسية، وهي ت t, ث θ, د d, ذ ð, ر r, ز z, س s, ش š, ص S, ض D, ط T, ظ Ď, ل l, ن n. على سبيل المثال، كلمة الشمس *Al+šams* تنطق /’aššams/ وليست

---

١ يتم اختصار الصائت غير المشدد في نهاية كلمة عصا *ʿaSaA*.

٢ بتصنيف آخر، فإن جميع هذه الصوامت تاجية، بمعنى يتم نطقها بالجزء المرن من مقدمة الفم [٣٩]. باستثناء /j/ ويرجع السبب في كثير من الأحيان إلى أن الفونيم (فيما قبل العربية الكلاسيكية) تعتبر حنكية (غير تاجية) [٣٨] أو وقف طبقي مجهور (/g/) [٣٩]. والحالة في اللهجات العربية هي نفسها في العربية المعاصرة مع وجود بعض الاختلافات [٣٩].

/*'alšams/[1]. بقية الصوامت تسمى حروفاً قمرية، ولا تدغم بأل التعريف. على سبيل المثال، كلمة القمر *Al+qamar* تنطق /'alqamar/ وليس /*'aqqamar/. تتطلب قواعد الإملاء العربية إضافة الشدة على الحرف الشمسي لبيان الإدغام بدون حذف اللام المدغمة، مثال ذلك الشَّمس *Alš~ams*.

• **التنـــوين**: إن إملاء المورفيم النكرة المشكَّلة هو مثال آخر على الإملاء الصرف- صوتي (morpho-phonemic spelling) الذي ذُكر في نقاشنا عن التشكيل فيما سبق.

• **الحروف الصـــامتة**: تظهر الألف الصامتة مع مورفيم (واو الجماعة) +وا /ū/ +*uwA*، وتشير إلى جمع مذكر مقترن بالأفعال. كما تظهر ألف صامتة أخرى في آخر الكلمة مع بعض الأسماء المنونة (قبل التشكيل أو بعده)، مثال ذلك، كتابا *kitaAbAã* أو *kitaAbãA* تنطق /kitāban/. وفي بعض القراءات الشعرية، يمكن معاملة الألف على شكل صائت طويل /ā/ ليصبح /kitābā/. وأخيرًا، هناك طريقة إملاء غريبة وشائعة لاسم العلم عمرو *ςamrw* وتنطق /ςamr/ حيث إن الواو (و *w*) في نهاية الاسم صامتة.

### ٣,٢,٤ قضايا التقييس

تم تقييس (standardization) التهجئة في اللغة العربية المعاصرة منذ فترة طويلة. ومع ذلك، لا تزال هناك بعض الاختلافات عبر البلدان وداخل البلدان العربية المختلفة. على سبيل المثال، هناك طريقتان شائعتان لإملاء بعض أسماء أعلام المناطق الجغرافية المنتهية بالصائت /a/ مثل /sūrya/ تظهر سوريا *swryA* أو سورية *swryħ*،

---

١ علامة النجمة (*) التي تسبق المثال هي علامة لغوية تعني أن المثال غير صحيح. وليس للنجمة علاقة برمز النجمة المستخدمة في الأنماط التعبيرية (regular expressions)، أو النقل الكتابي في بكوالتر للحرف ذ *ð*، أو حتى للنجمة المستخدمة لبيان التحليل في السياق المختار في بنك بنسلفانيا للتحاليل النحوية.

كذلك /'afrīqya/ تظهر أفريقيا *ÂfryqyA* و أفريقية *Âfryqyħ*. كما أن إملاء الهمزة فيها بعض الاستثناءات أيضا. على سبيل المثال ، /mas'ūl/ تظهر مسؤول *mswŵl* (شائعة في الشامية) و مسئول *msyŵl* (شائعة بالمصرية). أمثلة أخرى تتضمن إملاء الصوائت في الكلمات المقترضة (loan words) ، مثال ذلك أروبا *ÂrwbA* أو أوروبا *ÂwrwbA* (كلاهما ينطقان /'urubba/ مع التشديد على الصائت الثاني) ، وفلم *flm* أو فيلم *fylm* (تنطق /film/). أما بالنسبة للهجات العربية ، فلا توجد قواعد هجائية موحدة. ونتيجة لذلك ، لم يكن هناك قدر من التطابق في كتابة اللهجات كما هو في العربية المعاصرة.

**سؤال شائع: ما مدى صحة الادعاء بأنه "لا يوجد في العربية صوائت"؟**

هذه عبارة شائعة تذكر عن اللغة العربية ، وهو خطأ مضحك. وهناك عبارة أخرى أضعف في الادعاء تقول إن "**اللغة العربية لا تكتب الصوائت**" ، ومثال ذلك لقراء الإنجليزية هو ما يقوم به العرب عند قراءة جملة مثل ths s wht n rbc txt lks lk wth n vwls (وهي جملة منزوعة الصوائت من هذه الجملة " this is what an Arabic text looks like with no vowels "). هذا الادعاء أيضاً غير صحيح. كما شاهدنا ، حتى مع حذف جميع التشكيل الاختياري ، فلا تزال التهجئة العربية تعرض بعض الصوائت كالتالي : (أ) جميع الصوائت الطويلة تمثل بالحروف *A,w, y*. (ب) جميع الصوائت الاستهلالية يشار إليها بألف لوحدها ا A ، (ج) بعض الصوائت النهائية يشار إليها برمز إملائي صرف- صوتي مثل التاء المربوطة أو الألف المقصورة. وهناك تشبيه أكثر ملائمة للجملة الإنجليزية السابقة بعد حذف الصوائت منها *ths is wht an arbc txt lks lik wth no vwls*. واللافت للانتباه هنا أن الجملة السابقة تبدو مثل الكثير من الرسائل النصية الإنجليزية. ولربما استخدمت طرق التشكيل وفك الغموض في العربية لإرجاع الصوائت في الرسائل النصية الإنجليزية.

## ٣,٣ مهام معالجة اللغة الطبيعية

### ٣,٣,١ النقل الكتابي لأسماء الأعلام

النقل الكتابي لأسماء الأعلام (proper name transliteration) مشكلة فرعية محددة في الترجمة الآلية تركز على مطابقة/ تقريب القيمة الصوتية لأسماء الأعلام من لغة لأخرى وعادة عبر الخطوط (across scripts). في سياق النقل من العربية للإنجليزية والعكس، نواجه بعض التحديات منها:

- التشكيل الاختياري،
- الصوامت العربية التي ليس لها مطابق في الخط اللاتيني مثل /H/ و /ç/،
- اختلاف اللهجات في نطق الأسماء العربية،
- الصوامت الإنجليزية الغريبة على العربية مثل /p/ (تقرب إلى ب /b/) و /v/ (تقرب إلى ف /f/).
- الأسماء بالإنجليزية التي مصدرها لغات أوروبية أخرى تكتب أيضاً بالخط اللاتيني وتحضر معها التحديات الهجائية والصوتية الخاصة بها، مثل نطق الفرنسية التي تحذف الصوامت في النهاية بالإضافة إلى الطرق المختلفة لكتابة نفس الفونيم، مثل /š/ تكتب: ch , sh , sch وغيرها.

هناك العديد من التمثيلات لمشكلة النقل الكتابي لأسماء الأعلام. نستعرض هنا بعض الأمثلة:

- تشير مشكلة القذافي (Qaddafi problem) إلى الحالة التي يكون فيها إملاء معين في العربية يقابل عدة أشكال إملائية في الإنجليزية. بينما يملى اسم الرئيس الليبي بالعربية إلى قذافي (*qað~Afiy*)، يتم إملاء الاسم بالإنجليزية إلى التالي: Qadafi, Qaddafi, Gaddafi, Kaddafi, Kadafy.

- تشير مشكلة شوارزينغر (Schwarzenegger problem) إلى الحالة التي يكون فيها إملاء اسم بالإنجليزية يقابل عدة طرق إملائية بالعربية. هنا، يظهر الإملاء الصحيح الوحيد لاسم حاكم ولاية كاليفورنيا في اللغة العربية بالأشكال التالية: شوارزنيغر *šwArznyγr* ، شوارزنغر *šwArznγr* ، شوارزنيجر *šwArznyjr* ، شوارتزنجر *šwArtznjr* وغيرها. والبديل لهذه المشكلة هي حالة موزارت ( Mozart case) وفيها تحتفظ طريقتان للإملاء بطريقة نطق خاصة بالعربية، مثالها: موزارت *mwzArt* (الإنجلوفونيكية) و موزار *mwzAr* (الفرنكوفونية).
- تشير مشكلة حسن (*Hassan problem*) إلى الحالات التي يكون فيها إملاء محدد بالعربية لأسماء مختلفة يحصل لها لبس في الإنجليزية. فالنقحرة للاسم (*Hassan*) قد يؤل إلى حسن /Hasan/ أو حسان /Hassān/. وحصل الغموض المضاف هنا نتيجة لعدم وجود طريقة للإشارة إلى الإدغام في إملاء الإنجليزية، وخاصة عندما يستخدم حرف s مكرراً لإجبار نطق /s/ (مقارنة بنطق /z/). مثال أكثر تعقيدًا يظهر في الاسم *Salem* الذي من الممكن أن يكون النقل الكتابي الإنجلوفوني للاسم العربي سالم هو /sālim/ *sAlim* والنقل الكتابي الفرنكفوني للاسم العربي سلام هو /salēm/ *salAm* (كما تنطق في تونس).
- تشير مشكلة ماري (*Mary/Mari/Marie*) للحالات التي تكون فيها إملاءات مختلفة بالإنجليزية لنفس الاسم تكتب بشكل واحد في العربية. في الغالب تظهر الأسماء الثلاثة (*Mary ,Mari, Marie*) بالإملاء العربي على الشكل التالي: ماري *mAry*. ويمكن أن يحدث هذا أيضاً في بعض الأسماء العربية الغامضة إملائياً، مثال ذلك: Salim, Seleem , Slim هي ثلاثة طرق إملاء بالإنجليزية لنفس الاسم التاريخي العربي "سليم *slym*" متأثراً بالنطق الشائع له في الشام ومصر والمغرب. بشكل أو

بآخر هذه المشكلة لها علاقة بمشكلة القذافي السابق ذكرها ماعدا أن طرق الإملاء المتنوعة هنا تعتبر مميزة في إشارتها لأفراد مختلفين.

- تشير مشكلة القدس (*Urshalim/Alquds*) للحالات التي لا يوجد لها مقابل صوتي أو أن تشابهها الصوتي جزئي. على سبيل المثال، الاسم العربي لكلمة (*Jerusalem*) هو القدس *Alquds*. الاسم العبري للمدينة بالعربية هو أورشليم *Awršlym* ، الذي يحمل الكثير من الشبه بكلمة (*Jerusalem*) وذلك لأن الاسم الإنجليزي قد أتى من العبرية.

من المهم أن نتذكر أن الأخطاء في النقل الكتابي للاسم يمكن أن يكون لها تبعات كبيرة على حياة حاملي الاسم، على سبيل المثال، عن طريق الخلط ظلماً بينهم وبين الأفراد المشتبه بهم. وقد تلقت مشكلة النقحرة (الرومنة) لأسماء الأعلام الكثير من الاهتمام في مجال المعالجة الآلية للغة، وقد تناولت في نطاق واسع من الحلول في الأبحاث التالية [21، 43، 44، 45، 22].

### ٣,٣,٢ التصحيح الإملائي

ينظر إلى التصحيح الإملائي عادة على أنه خطوة تجهيزية (preprocessing step) تعالج وجود الأخطاء الإملائية. ويمكن أن تتسبب الأخطاء الإملائية في جعل نماذج معالجة اللغة الطبيعية أقل فاعلية، كما يمكنها أن تضيف هامش خطأ لا يمكن إصلاحه من الخطوة الأولى للنظام. ومن الصعب تحديد الأخطاء الإملائية إذا كانت الصياغة الإملائية للكلمة صحيحة لكن بوجود أخطاء صرفية أو لغوية [46].

كما ذكرنا في الفصل الثاني، أكثر الأخطاء الإملائية في العربية تتضمن اللبس في الألف المهموزة والألف/الياء المقصورة. تؤثر هذه الأخطاء في ١١٪ من جميع

الكلمات (أو ٤,٥ من الأخطاء لكل جملة) في بنك بنسلفانيا للتحاليل النحوية [٤٧]. شكل آخر من الأخطاء يتضمن وضع النقاط في غير محلها، بينما الكلمات المجمعة أو المفرقة بالقرب من الحروف المنفصلة، أو الحروف في غير محلها تحدث بتكرار أقل – ٠,٣٪ من جميع الكلمات (على الأقل واحدة من حوالي ١٢٪ من الجمل). ولا تزال هذه النسبة غير مستهان بها، حيث إن خطأ إملائياً واحداً يمكن أن يعيث فساداً في معالجة الجملة بأكملها.

تتضمن الأبحاث [48، 49، 46، 50] أمثلة على جهود موجهة لتصحيح الإملاء آلياً أو معالجة الأخطاء الإملائية لتطبيقات معالجة اللغة آلياً. والتوليد الآلي لبدائل أشكال الألف المهموزة والألف المقصورة الصحيحة عن طريق بعض المحللات الصرفية خارج السياق، مثال ذلك [23]. وتختار أداة المحلل الصرفي ومزيل الغموض العربي آلياً (مدى MADA) [51] الألف المهموزة والألف المقصورة المناسبة في السياق كجزء من منهجية إزالة الغموض الصرفي العام. وقد ذكر أنها ناجحة بنسبة ٩٩,٤٪ في أداء هذه المهمة [13].

بما أن اللهجات العربية غير مقيسة، فإنه من المهم أن نشير إلى أن تهجئتها ليست دائماً ثابتة. وعلى هذا النحو، يحتاج استخدام أي طريقة لمعالجة اللهجات إلى اتفاقية داخلية لاستخدام إملاء معياري [52، 53، 54].

### ٣,٣,٣ التعرف على الكلام وتوليده آلياً

التعرف على الكلام (Speech Recognition) أيضاً يعرف بالتعرف الآلي على الكلام (Automatic Speech (Recognition – ASR;) أو تحويل الصوت لنص (Speech-to-Text – STT) هي عملية تحويل إشارة الكلام الصوتية إلى سلسلة من الكلمات

المقابلة لها. في المقابل، توليد الكلام آلياً (Speech Synthesis) أو ما يعرف بتحويل النص لصوت (Text-to-Speech –TTS ) هي عملية إنتاج إشارة صوتية من نص مدخل. وقد تم إجراء الكثير من البحوث في كلا المجالين [55، 56، 33، 57، 58، 41، 59، 41]. أيضاً هناك الكثير من الموارد لتدريب واختبار النظام (انظر الملحق ج).

لا تعتمد معظم الطرق المستخدمة للتعرف الآلي على الكلام (ASR) وتوليد الكلام آلياً (TTS) على اللغة، وذلك عند تحديد مستوى التمثيل المناسب للغة المعنية. والتحدي الكبير بالنسبة للغة العربية، هو سد الفجوة بين الأصوات العربية وتهجئتها، حيث إنه عادة إذا كان الإملاء معقَّداً وركيكاً، فإن اللغة تكون صعبة على أنظمة التعرف على الكلام (ASR) أو توليده (TTS).

وعلى هذا النحو، فإن المهمة المركزية للعمل على التعرف على الكلام وتوليده للغة العربية يشمل إنتاج الشكل الفونيمي أو الصوتي للنص العربي. وقد لوحظ أن التشكيل لوحده لا يمكنه التنبؤ بالنطق الفعلي للغة العربية المعاصرة [41]. وقد وصف باحثون مختلفون مجموعات مختلفة من قواعد النطق بناء على علم الأصوات في اللغة العربية المعاصرة حيث يطبق نطق كلمة تم تشكيلها على مجموعة من طرق النطق الممكنة. وتحاول بعض من هذه القواعد استيعاب متغيرات النطق للتعامل مع الإخفاقات الشائعة وذلك لإنتاج النطق "المناسب" وفقاً للنحو العربي وعلم الأصوات العربية حتى من قبل متحدثين مدربين على اللغة العربية المعاصرة [42، 58، 41].

والتحدي الآخر هو الصرف العربي المعقد والغني، مما يؤدي إلى إنتاج مفردات كثيرة جداً لتغطيتها. ويلاحظ أن اللغة العربية لديها معدل نمو يبلغ ٢,٥ مقارنة بالمفردات في اللغة الإنجليزية ولديها أيضا ١٠ أضعاف المفردات غير المعروفة (out-of-vocabulary) (OOV) (في قاموس يحوي ٦٤ ألف كلمة) [60، 61]. يتم

معالجة هذه المسألة من خلال تقديم نماذج صرفية لتقليل معدلات حجم المفردات والمفردات غير المعروفة (OOV) [60، 61].

أما بالنسبة للاختلافات في اللهجات، فإن المتفق عليه في هذا المجال من قبل مطوري تقنيات اللغة هو أن نظام التعرف الآلي على الكلام ينبغي أن يكون متيناً بما فيه الكفاية للتعامل مع اللغة العربية المعاصرة، وأيضاً في الخلط بين الفصحى والعامية (العربية المعاصرة واللهجات). ومع ذلك، فإن عملية تحويل النص إلى كلام (TTS) يجب أن تركز فقط على إنتاج اللغة العربية المعاصرة. والاستثناء الوحيد هو العمل المنجز على الترجمة الآلية للكلام إلى كلام في اللغة الإنجليزية ⇔ العراقية في مشروع (TRANSTAC)[1] والممول من قبل (DARPA) [62].

### ٣,٤ المزيد من القراءات

هناك عدد متزايد من الأبحاث في مجال التحديد الآلي للهجة، حيث تتمثل المهمة في التعرف على اللهجات العربية من الإشارات الصوتية [63، 65، 64]. كما يوجد أيضاً بعض الأعمال المثيرة للاهتمام بشأن التحديد التلقائي للجوانب العاطفية في الخطاب (في العربية من بين اللغات الأخرى) وتحديداً الكاريزما وكيف ينظر إليها بشكل مختلف عبر الثقافات [66].

---

١ TRANSTAC هو اختصار (Spoken Language Communication and Translation System for Tactical Use) نظام للترجمة والتواصل للغة المنطوقة للاستخدام التكتيكي.